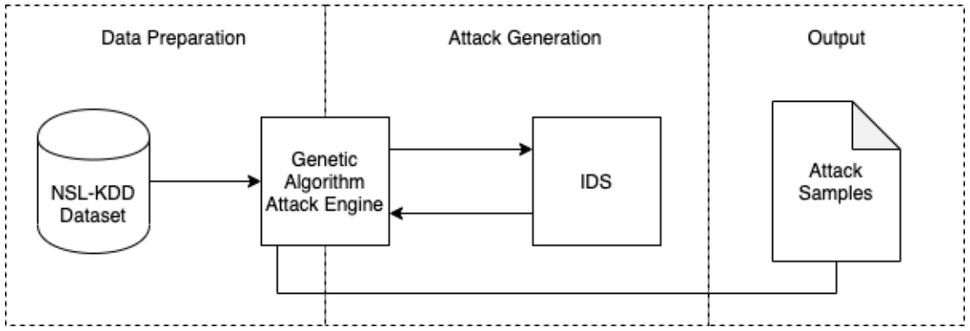
Evasion Attacks Against Intrusion Detection Systems Using Genetic Algorithms

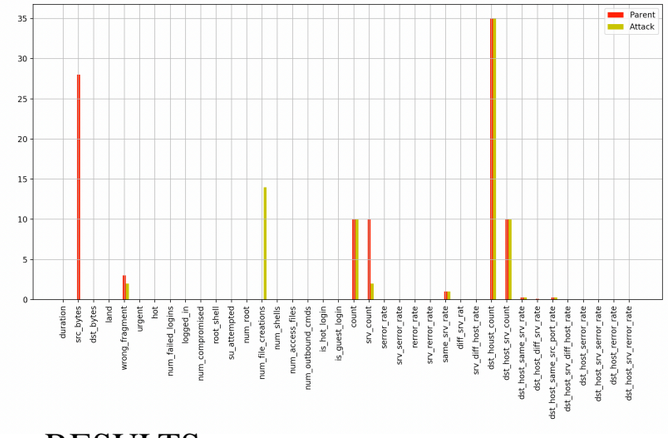
Raymond Mogg & Associate Professor Dan Kim

AIM

"generate interpretable evasion attacks against an intrusion detection systems using genetic algorithms. The IDS will be network based, and use a behaviour based detection method"

ATTACK PIPELINE





BACKGROUND

Intrusion Detection Systems

Intrusion Detection Systems (IDS) are used to monitor a network in order to detect attempted malicious attacks against it.

Genetic Algorithms

Genetic algorithms are a family of machine learning models based on evolution, often used on optimisation problems.

Evasion Attacks

Evasion attacks are an attack against an intrusion detection system where a malicious actor crafts an input in order to get an attack sample to be classified as benign. Essentially it is crafted misclassification.

DATA PREPARATION

Cleaning and Preprocessing

Data from the NSL-KDD dataset is processed into sklearn and cleaned. A attack type to focus on is defined in the algorithm setup.

Decision Tree Generation

A decision tree is trained on the cleaned data in order to produce an IDS for attack generation.

Attack Seeding

A randomly selected attack of the type being investigated is selected from the dataset. This is used as part of the genetic algorithm as a base level attack for comparison purposes.

ATTACK GENERATION

Genetic Algorithm

The genetic algorithm is used in order to generate attack samples. By using a feedback loop between the algorithm and the IDS, the fitness of each sample can be evaluated and is used to selected which offspring survive and which offspring die in each generation, helping the best samples progress forward.

This process runs for 20 iterations, with each iteration producing 120 offspring. The fittest 30 go through to the next generation.

ATTACK OUTPUT

Final Attacks

At the conclusion of the algorithm, 30 final samples are generated. These final samples are arrays of network connection data in the same form as the original NSL-KDD input. Each sample is then run through the IDS and can be checked if it is classified as an attack or benign. Generating benign samples with a similar structure to the original attack sample is the optimal goal for the algorithm. Each iteration commonly produces between 10 and 20 attack samples.

RESULTS

Generated Attacks

The attack pipeline can successfully generate sample attacks that are similar in content to an attack, but are classified as benign by the IDS.

The produced samples have characteristics similar to the decision boundary of the IDS (its feature variable values were close to decision boundary values). This indicates the pipeline could be generalised to other types of IDSes, not just decision trees, and it should be able to still produce attack samples.

Future Work - Validation of Results

Currently sample attacks are not validated against a real system. Validation of attacks would prove more insight into whether the attacks generated are actually attacks or in fact benign samples still.